Capítulo 3

Métodos Jerárquicos de Análisis Cluster.

3.1. Introducción.

Los llamados métodos jerárquicos tienen por objetivo agrupar clusters para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen en aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

1. Los métodos aglomerativos, también conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo conglomerado.

2. Los métodos disociativos, también llamados descendentes, constituyen el proceso inverso al anterior. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez más pequeños. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

Para fijar ideas, centrémonos un segundo en los métodos aglomerativos. Sea n el conjunto de individuos de la muestra, de donde resulta el nivel $K = 0$, con n grupos. En el siguiente nivel se agruparán aquellos dos individuos que tengan la mayor similitud (o menor distancia), resultando así $n - 1$ grupos; a continuación, y siguiendo con la misma estrategia, se agruparán en el nivel posterior, aquellos dos individuos (o clusters ya formados) con menor distancia o mayor similitud; de esta forma, en el nivel L tendremos $n - L$ grupos formados. Si se continúa agrupando de esta forma, se llega al nivel $L = n - 1$ en el que sólo hay un grupo, formado por todos los individuos de la muestra.

Esta manera de formar nuevos grupos tiene la particularidad de que si en un determinado nivel se agrupan dos clusters, éstos quedan ya jerárquicamente agrupados para el resto de los niveles.

Los métodos jerárquicos permiten la construcción de un árbol de clasificación, que recibe el nombre de dendrograma (figura 3.1), en el cual se puede seguir de forma gráfica el procedimiento de unión seguido, mostrando que grupos se van uniendo, en que nivel concreto lo hacen, así como el valor de la medida de asociación entre los grupos cuando éstos se agrupan (valor que llamaremos nivel de fusión).

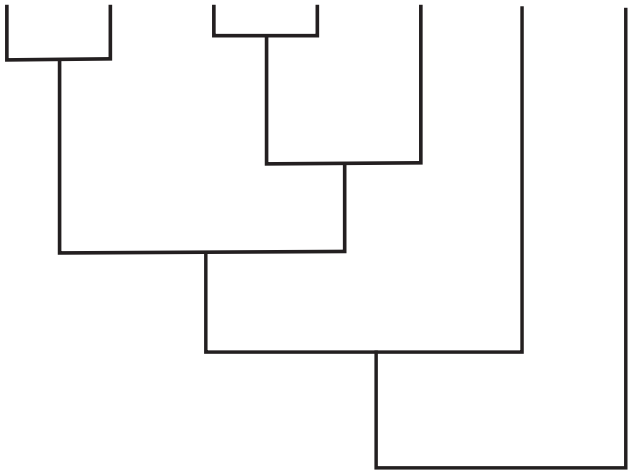
En resumen, la forma general de operar de estos métodos es bastante simple. Por ejemplo, en los métodos aglomerativos se parte de tantos grupos como individuos haya. A continuación se selecciona una medida de similitud, agrupándose los dos grupos o clusters con mayor similitud. Así se continúa hasta que:

1. Se forma un solo grupo.

2. Se alcanza el número de grupos prefijado.

3. Se detecta, a través de un contraste de significación, que hay razones estadísticas para no continuar agrupando clusters, ya que los más similares no son lo suficientemente homogéneos como para determinar una misma agrupación.

Figura 3.1: Dendrograma



3.2. Métodos Jerárquicos Aglomerativos.

A continuación vamos a presentar algunas de las estrategias que pueden ser empleadas a la hora de unir los clusters en las diversas etapas o niveles de un procedimiento jerárquico. Ninguno de estos procedimientos proporciona una solución óptima para todos los problemas que se pueden plantear, ya que es
posible llegar a distintos resultados según el método elegido. El buen criterio del investigador, el conocimiento del problema planteado y la experiencia, sugerirán el método más adecuado. De todas formas, es conveniente, siempre, usar varios procedimientos con la idea de contrastar los resultados obtenidos y sacar conclusiones, tanto como si hubiera coincidencias en los resultados obtenidos con métodos distintos como si no las hubiera.

3.2.1. Estrategia de la distancia mínima o similitud máxima.

Esta estrategia recibe en la literatura anglosajona el nombre de *amalgamamiento simple (single linkage)*. En este método se considera que la distancia o similitud entre dos clusters viene dada, respectivamente, por la mínima distancia (o máxima similitud) entre sus componentes.

Así, si tras efectuar la etapa K-ésima, tenemos ya formados $n-K$ clusters, la distancia entre los clusters C_i (con n_i elementos) y C_j (con n_j elementos) sería:

$$d(C_i, C_j) = \min_{\substack{x_l \in C_i \\ x_m \in C_j}} \{d(x_l, x_m)\} \quad l = 1, \ldots, n_i \ ; \ m = 1, \ldots, n_j \tag{3.1}$$

mientras que la similitud, si estuviéramos empleando una medida de tal tipo, entre los dos clusters sería:

$$s(C_i, C_j) = \max_{\substack{x_l \in C_i \\ x_m \in C_j}} \{s(x_l, x_m)\} \quad l = 1, \ldots, n_i \ ; \ m = 1, \ldots, n_j \tag{3.2}$$

Con ello, la estrategia seguida en el nivel $K+1$ será:

1. En el caso de emplear distancias, se unirán los clusters C_i y C_j si

$$d(C_i, C_j) = \min_{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}} \{d(C_{i_1}, C_{j_1})\} =$$

$$= \min_{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}} \left\{ \min_{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}} \{d(x_l, x_m)\} \right\} \quad l = 1, \ldots, n_{i_1} \ ; \ m = 1, \ldots, n_{j_1}$$

2. En el caso de emplear similitudes, se unirán los clusters C_i y C_j si

$$s(C_i, C_j) = \max_{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}} \{s(C_{i_1}, C_{j_1})\} =$$

$$= \underset{\substack{i_1,j_1=1,\ldots,n-K \\ i_1 \neq j_1}}{\text{Max}} \left\{ \underset{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}}{\text{Max}} \{s(x_l, x_m)\} \right\} \quad l = 1, \ldots, n_{i_1} \ ; \ m = 1, \ldots, n_{j_1}$$

donde, como es natural, se ha seguido la norma general de maximizar las similitudes o bien minimizar las distancias.

Ejemplo 3.1 *Partiendo de la matriz de distancias inicial entre 7 individuos*

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	2,15	0					
C	0,7	1,53	0				
D	1,07	1,14	0,43	0			
E	0,85	1,38	0,21	0,29	0		
F	1,16	1,01	0,55	0,22	0,41	0	
G	1,56	2,83	1,86	2,04	2,02	2,05	0

los pasos seguidos en un procedimiento cluster jerárquico ascendente, empleando la estrategia del amalgamamiento simple, serían los siguientes:

1. **Nivel K=1**

   $\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$*, por lo que el primer cluster que se forma es el cluster* (C, E)*.*

2. **Nivel K=2**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,7	1,38	0			
D	1,07	1,14	0,29	0		
F	1,16	1,01	0,41	0,22	0	
G	1,56	2,83	1,86	2,04	2,05	0

   Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$*, por lo que se forma el cluster* (D, F)*.*

3. **Nivel K=3**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,7	1,38	0		
(D,F)	1,07	1,01	0,29	0	
G	1,56	2,83	1,86	2,04	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}29$*, formándose el cluster* $((C, E), (D, F))$*.*

4. **Nivel K=4**

   La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	0,7	1,01	0	
G	1,56	2,83	1,86	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 0{,}7$*, formándose el cluster* $(A, ((C, E), (D, F)))$*.*

5. **Nivel K=5**

La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	1,01	0	
G	1,56	2,83	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{B, (A, ((C, E), (D, F)))\} = 1{,}01$*, formándose el cluster* $(B, (A, ((C, E), (D, F))))$*.*

6. **Nivel K=6**

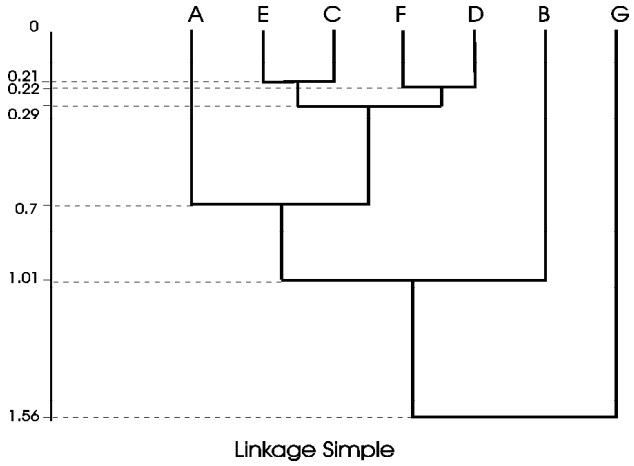
La matriz de distancias en este paso es:

	(B,(A,((C,E),(D,F))))	**G**
(B,(A,((C,E),(D,F))))	0	
G	1,56	0

Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado es el de la figura 3.2

Figura 3.2: Método del amalgamamiento simple



3.2.2. Estrategia de la distancia máxima o similitud mínima.

En este método, también conocido como el procedimiento de *amalgamamiento completo (complete linkage)*, se considera que la distancia o similitud entre dos clusters hay que medirla atendiendo a sus elementos más dispares, o sea, la distancia o similitud entre clusters viene dada, respectivamente, por la máxima distancia (o mínima similitud) entre sus componentes.

Así pues, al igual que en la estrategia anterior, si estamos ya en la etapa K-ésima, y por lo tanto hay ya formados $n - K$ clusters, la distancia y similitud entre los clusters C_i y C_j (con n_i y n_j elementos respectivamente), serán:

$$d(C_i, C_j) = \max_{\substack{x_l \in C_i \\ x_m \in C_j}} \{d(x_l, x_m)\} \quad l = 1, \ldots, n_i \ ; \ m = 1, \ldots, n_j \tag{3.3}$$

$$s(C_i, C_j) = \min_{\substack{x_l \in C_i \\ x_m \in C_j}} \{s(x_l, x_m)\} \quad l = 1, \ldots, n_i \ ; \ m = 1, \ldots, n_j \tag{3.4}$$

y con ello, la estrategia seguida en el siguiente nivel, $K + 1$, será:

1. En el caso de emplear distancias, se unirán los clusters C_i y C_j si

$$d(C_i, C_j) = \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Min}} \{d(C_{i_1}, C_{j_1})\} =$$

$$= \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Min}} \left\{ \underset{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}}{\text{Max}} \{d(x_l, x_m)\} \right\} \quad l = 1, \ldots, n_{i_1} \ ; \ m = 1, \ldots, n_{j_1}$$

2. En el caso de emplear similitudes, se unirán los clusters C_i y C_j si

$$s(C_i, C_j) = \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Max}} \{s(C_{i_1}, C_{j_1})\} =$$

$$= \underset{\substack{i_1, j_1 = 1, \ldots, n-K \\ i_1 \neq j_1}}{\text{Max}} \left\{ \underset{\substack{x_l \in C_{i_1} \\ x_m \in C_{j_1}}}{\text{Min}} \{s(x_l, x_m)\} \right\} \quad l = 1, \ldots, n_{i_1} \ ; \ m = 1, \ldots, n_{j_1}$$

Ejemplo 3.2 *En el mismo ejemplo anterior se tendrá:*

1. **Nivel K=1**

$\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$*, por lo que el primer cluster que se forma es el cluster* (C, E)*.*

2. **Nivel K=2**

La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,85	1,53	0			
D	1,07	1,14	0,43	0		
F	1,16	1,01	0,55	0,22	0	
G	1,56	2,83	2,02	2,04	2,05	0

Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$*, por lo que se forma el cluster* (D, F)*.*

3. **Nivel K=3**

La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,85	1,53	0		
(D,F)	1,16	1,14	0,55	0	
G	1,56	2,83	2,02	2,05	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}55$*, formándose el cluster* $((C, E), (D, F))$*.*

4. **Nivel K=4**

La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	1,16	1,53	0	
G	1,56	2,83	2,05	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 1{,}16$*,*
formándose el cluster $(A, ((C, E), (D, F)))$*.*

5. **Nivel K=5**

La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	2,15	0	
G	2,05	2,83	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{G, (A, ((C, E), (D, F)))\} = 2{,}05$*, formándose el cluster* $(G, (A, ((C, E), (D, F))))$.

6. **Nivel K=6**

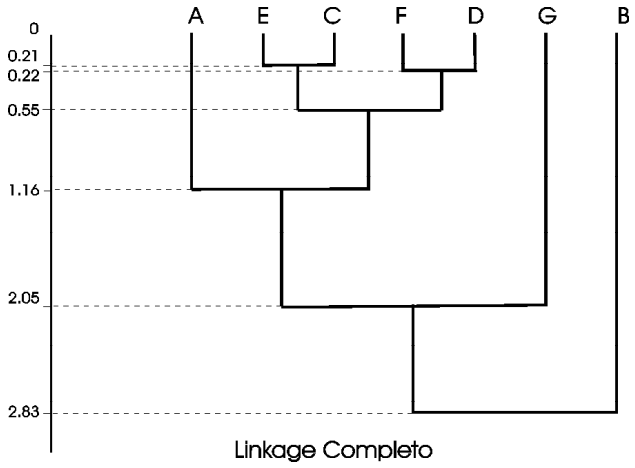
La matriz de distancias en este paso es:

	(G,(A,((C,E),(D,F))))	**B**
(G,(A,((C,E),(D,F))))	0	
B	2,83	0

Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado es el de la figura 3.3

Figura 3.3: Método del amalgamamiento completo



3.2.3. Estrategia de la distancia, o similitud, promedio no ponderada. (Weighted aritmethic average)

En esta estrategia la distancia, o similitud, del cluster C_i con el C_j se obtiene como la media aritmética entre la distancia, o similitud, de las componentes de dichos clusters.

Así, si el cluster C_i (con n_i elementos) está compuesto, a su vez, por dos clusters C_{i_1} y C_{i_2} (con n_{i_1} y n_{i_2} elementos respectivamente), y el cluster C_j posee n_j elementos, la distancia, o similitud, entre ellos se calcula como

$$d(C_i, C_j) = \frac{d(C_{i_1}, C_j) + d(C_{i_2}, C_j)}{2} \tag{3.5}$$

Notemos que en este método no se tiene en cuenta el tamaño de ninguno de los clusters involucrados en el cálculo, lo cual significa que concede igual importancia a la distancia $d(C_{i_1}, C_j)$ que a la distancia $d(C_{i_2}, C_j)$.

Ejemplo 3.3 *Continuando con el ejemplo anterior, ahora tendremos:*

1. **Nivel K=1**

   $\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$, *por lo que el primer cluster que se forma es el cluster* (C, E).

2. **Nivel K=2**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,775	1,455	0			
D	1,07	1,14	0,36	0		
F	1,16	1,01	0,48	0,22	0	
G	1,56	2,83	1,94	2,04	2,05	0

   Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$, *por lo que se forma el cluster* (D, F).

3. **Nivel K=3**

   La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,775	1,455	0		
(D,F)	1,115	1,075	0,42	0	
G	1,56	2,83	1,94	2,045	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}42$, *formándose el cluster* $((C, E), (D, F))$.

4. **Nivel K=4**

   La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	0,945	1,265	0	
G	1,56	2,83	1,9925	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 0{,}945$, *formándose el cluster* $(A, ((C, E), (D, F)))$.

5. **Nivel K=5**

   La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	1,7075	0	
G	1,77625	2,83	0

   En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{B, (A, ((C, E), (D, F)))\} = 1{,}7075$, *formándose el cluster* $(B, (A, ((C, E), (D, F))))$.

6. **Nivel K=6**

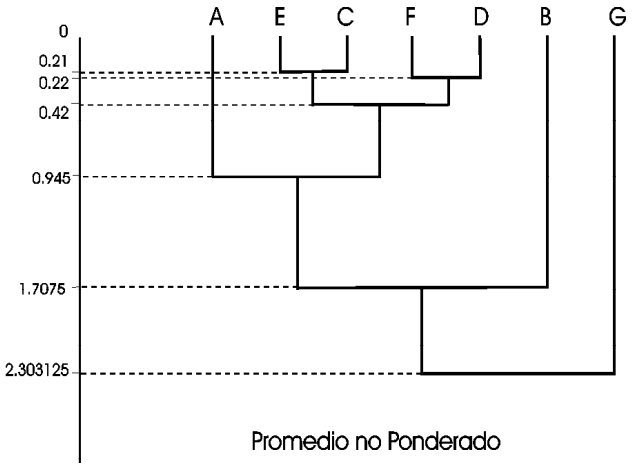
   La matriz de distancias en este paso es:

	(B,(A,((C,E),(D,F))))	**G**
(B,(A,((C,E),(D,F))))	0	
G	2,303125	0

   Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado a este ejemplo es el de la figura 3.4

Figura 3.4: Método del promedio no ponderado



3.2.4. Estrategia de la distancia, o similitud, promedio ponderada. (unweighted aritmethic average)

Se considera que la distancia, o similitud, entre dos clusters, viene definida por el promedio ponderado de las distancias, o similitudes, de los componentes de un cluster respecto a los del otro.

Sea dos clusters, C_i y C_j; supongamos que el cluster C_i está formado, a su vez, por otros dos clusters, C_{i_1} y C_{i_2}, con n_{i_1} y n_{i_2} elementos respectivamente. Sea $n_i = n_{i_1} + n_{i_2}$ el número de elementos de C_i y n_j el número de elementos que componen C_j. Entonces, en términos de distancias (igual puede hacerse para similitudes), la distancia promedio ponderada sería, notando $x_i \in C_i$, $x_{i_1} \in C_{i_1}$, $x_{i_2} \in C_{i_2}$, $x_j \in C_j$

$$d(C_i, C_j) = \frac{1}{(n_{i_1} + n_{i_2})n_j} \sum_{i=1}^{n_{i_1}+n_{i_2}} \sum_{j=1}^{n_j} d(x_i, x_j) =$$

$$= \frac{1}{(n_{i_1} + n_{i_2})n_j} \sum_{i_1=1}^{n_{i_1}} \sum_{j=1}^{n_j} d(x_{i_1}, x_j) + \frac{1}{(n_{i_1} + n_{i_2})n_j} \sum_{i_2=1}^{n_{i_2}} \sum_{j=1}^{n_j} d(x_{i_2}, x_j) =$$

$$= \frac{n_{i_1}}{(n_{i_1} + n_{i_2})n_{i_1} n_j} \sum_{i_1=1}^{n_{i_1}} \sum_{j=1}^{n_j} d(x_{i_1}, x_j) + \frac{n_{i_2}}{(n_{i_1} + n_{i_2})n_{i_2} n_j} \sum_{i_2=1}^{n_{i_2}} \sum_{j=1}^{n_j} d(x_{i_2}, x_j) =$$

$$= \frac{n_{i_1}}{n_{i_1} + n_{i_2}} d(C_{i_1}, C_j) + \frac{n_{i_2}}{n_{i_1} + n_{i_2}} d(C_{i_2}, C_j) =$$

$$= \frac{n_{i_1} d(C_{i_1}, C_j) + n_{i_2} d(C_{i_2}, C_j)}{n_{i_1} + n_{i_2}} \tag{3.6}$$

con lo cual la distancia $d(C_i, C_j)$ es el promedio ponderado de las distancias de cada uno de los dos clusters previos, C_{i_1} y C_{i_2}, con respecto al cluster C_j.

Ejercicio 3.1 *Comprobar que, con la estrategia de la distancia promedio ponderada, se tiene*

$$d([(a,b),c],[(d,e),f]) = \frac{2d\left((d,e),[(a,b),c]\right) + d\left(f,[(a,b),c]\right)}{3} =$$

$$= \frac{d(a,d) + d(a,e) + d(a,f) + d(b,d) + d(b,e) + d(b,f) + d(c,d) + d(c,e) + d(c,f)}{9}$$

Ejemplo 3.4 *Siguiendo con el ejemplo tratado anteriormente, ahora tendremos:*

1. **Nivel K=1**

   $\text{Min}\{d(C_i, C_j)\} = d(C, E) = 0{,}21$*, por lo que el primer cluster que se forma es el cluster* (C, E).

2. **Nivel K=2**

La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**D**	**F**	**G**
A	0					
B	2,15	0				
(C,E)	0,775	1,455	0			
D	1,07	1,14	0,36	0		
F	1,16	1,01	0,48	0,22	0	
G	1,56	2,83	1,94	2,04	2,05	0

Ahora bien, $\text{Min}\{d(C_i, C_j)\} = d(D, F) = 0{,}22$*, por lo que se forma el cluster* (D, F)*.*

3. **Nivel K=3**

La matriz de distancias en este paso es:

	A	**B**	**(C,E)**	**(D,F)**	**G**
A	0				
B	2,15	0			
(C,E)	0,775	1,455	0		
(D,F)	1,115	1,075	0,42	0	
G	1,56	2,83	1,94	2,045	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{(C, E), (D, F)\} = 0{,}42$*, formándose el cluster* $((C, E), (D, F))$*.*

4. **Nivel K=4**

La matriz de distancias en este paso es:

	A	**B**	**((C,E),(D,F))**	**G**
A	0			
B	2,15	0		
((C,E),(D,F))	0,945	1,265	0	
G	1,56	2,83	1,9925	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{A, ((C, E), (D, F))\} = 0{,}945$*, formándose el cluster* $(A, ((C, E), (D, F)))$*.*

5. **Nivel K=5**

La matriz de distancias en este paso es:

	(A,((C,E),(D,F)))	**B**	**G**
(A,((C,E),(D,F)))	0		
B	1,442	0	
G	1,906	2,83	0

En este caso, $\text{Min}\{d(C_i, C_j)\} = d\{B, (A, ((C, E), (D, F)))\} = 1{,}442$*, formándose el cluster* $(B, (A, ((C, E), (D, F))))$*.*

6. **Nivel K=6**

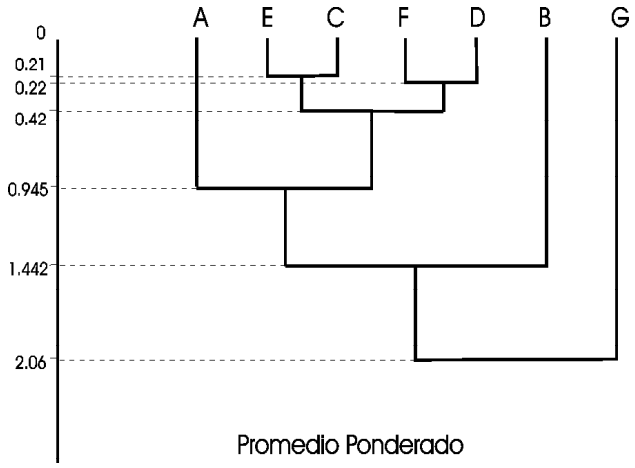
La matriz de distancias en este paso es:

	(B,(A,((C,E),(D,F))))	**G**
(B,(A,((C,E),(D,F))))	0	
G	2,06	0

Este será el último paso, en el cual, evidentemente, se tendrá un único cluster formado por los 7 individuos.

El dendrograma asociado a este ejemplo es el de la figura 3.5

Figura 3.5: Método del promedio ponderado



3.2.5. Métodos basados en el centroide.

En estos métodos, la semejanza entre dos clusters viene dada por la semejanza entre sus centroides, esto es, los vectores de medias de las variables medidas sobre los individuos del cluster.

Entre ellos distinguiremos dos:

1. Método del centroide ponderado, en el que los tamaños de los clusters son considerados a la hora de efectuar los cálculos.
2. Método del centroide no ponderado, o método de la mediana, en el cual los tamaños de los clusters no son considerados.

Veamos cada uno de ellos por separado:

1. En cuanto al primero de ellos y centrándonos en la distancia euclídea al cuadrado, supongamos que pretendemos medir la distancia entre los clusters C_j (compuesto por n_j elementos) y C_i (formado a su vez por dos clusters, C_{i_1} y C_{i_2}, con n_{i_1} y n_{i_2} elementos, respectivamente). Sean m^j, m^{i_1} y m^{i_2} los centroides de los clusters anteriormente citados (obviamente, esos centroides son vectores n dimensionales).

   Así, el centroide del cluster C_i vendrá dado en notación vectorial por:

$$m^i = \frac{n_{i_1} m^{i_1} + n_{i_2} m^{i_2}}{n_{i_1} + n_{i_2}}$$

   cuyas componentes serán:

$$m_l^i = \frac{n_{i_1} m_l^{i_1} + n_{i_2} m_l^{i_2}}{n_{i_1} + n_{i_2}} \quad l = 1, \ldots, n$$

   Con ello, la distancia euclídea al cuadrado entre los clusters C_i y C_j vendrá dada por:

$$d_2^2(C_j, C_i) = \sum_{l=1}^{n} \left(m_l^j - m_l^i \right)^2 = \sum_{l=1}^{n} \left[m_l^j - \frac{n_{i_1} m_l^{i_1} + n_{i_2} m_l^{i_2}}{n_{i_1} + n_{i_2}} \right]^2 =$$

$$= \sum_{l=1}^{n} \left[(m_l^j)^2 - 2 m_l^j \frac{n_{i_1} m_l^{i_1} + n_{i_2} m_l^{i_2}}{n_{i_1} + n_{i_2}} + \right.$$

$$\left. + \frac{(n_{i_1})^2 (m_l^{i_1})^2 + (n_{i_2})^2 (m_l^{i_2})^2 + n_{i_1} n_{i_2} (m_l^{i_1})^2 + n_{i_1} n_{i_2} (m_l^{i_2})^2}{(n_{i_1} + n_{i_2})^2} + \right.$$

$$+\frac{-n_{i_1}n_{i_2}(m_l^{i_1})^2 - n_{i_1}n_{i_2}(m_l^{i_2})^2 + 2n_{i_1}n_{i_2}m_l^{i_1}m_l^{i_2}}{(n_{i_1}+n_{i_2})^2}\Bigg] =$$

$$= \sum_{l=1}^{n}\Bigg[(m_l^j)^2 - 2m_l^j\frac{n_{i_1}m_l^{i_1}+n_{i_2}m_l^{i_2}}{n_{i_1}+n_{i_2}} +$$

$$+\frac{(n_{i_1}+n_{i_2})n_{i_1}(m_l^{i_1})^2 + (n_{i_1}+n_{i_2})n_{i_2}(m_l^{i_2})^2}{(n_{i_1}+n_{i_2})^2} -$$

$$-\frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}\left[(m_l^{i_1})^2 + (m_l^{i_2})^2 - 2m_l^{i_1}m_l^{i_2}\right]\Bigg] =$$

$$= \sum_{l=1}^{n}\Bigg[\frac{n_{i_1}(m_l^j)^2 + n_{i_2}(m_l^j)^2}{n_{i_1}+n_{i_2}} + \frac{n_{i_1}(m_l^{i_1})^2}{n_{i_1}+n_{i_2}} + \frac{n_{i_2}(m_l^{i_2})^2}{n_{i_1}+n_{i_2}} -$$

$$-2m_l^j\frac{n_{i_1}m_l^{i_1}}{n_{i_1}+n_{i_2}} - 2m_l^j\frac{n_{i_2}m_l^{i_2}}{n_{i_1}+n_{i_2}} - \frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}\left[m_l^{i_1}-m_l^{i_2}\right]^2\Bigg] =$$

$$= \sum_{l=1}^{n}\Bigg[\frac{n_{i_1}}{n_{i_1}+n_{i_2}}\left[m_l^j - m_l^{i_1}\right]^2 + \frac{n_{i_2}}{n_{i_1}+n_{i_2}}\left[m_l^j - m_l^{i_2}\right]^2 -$$

$$-\frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}\left[m_l^{i_1}-m_l^{i_2}\right]^2\Bigg] =$$

$$= \frac{n_{i_1}}{n_{i_1}+n_{i_2}}d_2^2(C_{i_1},C_j) + \frac{n_{i_2}}{n_{i_1}+n_{i_2}}d_2^2(C_{i_2},C_j) - \frac{n_{i_1}n_{i_2}}{(n_{i_1}+n_{i_2})^2}d_2^2(C_{i_1},C_{i_2}) \tag{3.7}$$

Comentario 3.2.1 *Nótese que la relación anterior se ha establecido para el caso particular de la distancia euclídea. No obstante, dicha relación se sigue verificando si la distancia empleada viene definida a partir de una norma que proceda de un producto escalar.* [1]

Esta hipótesis puede relajarse aún más hasta considerar distancias que procedan de una norma que verifique la ley del paralelogramo

$$||x+y||^2 + ||x-y||^2 = 2\left[||x||^2 + ||y||^2\right]$$

ya que en tales circunstancias se puede definir un producto escalar a partir de ella como

$$< x, y >= \frac{1}{4}\left[||x+y||^2 - ||x-y||^2\right]$$

2. Una desventaja del procedimiento anterior estriba en que si los tamaños n_{i_1} y n_{i_2} de los componentes de C_i son muy diferentes entre sí, se corre el peligro de que el centroide de dicho cluster, m^i, esté influenciado excesivamente por el componente con tamaño superior y las cualidades del grupo pequeño no se tengan prácticamente en cuenta.

   Así la estrategia de la distancia mediana, al considerar de forma arbitraria que $n_{i_1} = n_{i_2}$, provoca que el centroide del cluster C_i esté situado entre los clusters C_{i_1} y C_{i_2} y con ello el centroide del cluster (C_i, C_j) esté localizado en el punto central o mediana del triángulo formado por los clusters C_{i_1}, C_{i_2} y C_j.

   Salvo esta diferencia, la estrategia de la distancia mediana es análoga a la anterior y, por lo tanto, goza de sus mismas características. Así, si estamos hablando de distancias, la distancia entre el cluster C_i y el C_j viene dada por

[1] Dado un producto escalar en un espacio vectorial, se puede definir la norma de un vector como la raíz cuadrada positiva del producto escalar del vector por sí mismo.

$$d(C_i, C_j) = \frac{1}{2}\left[d(C_{i_1}, C_j) + d(C_{i_2}, C_j)\right] - \frac{1}{4}d(C_{i_1}, C_{i_2})$$

y si hablamos de similitudes

$$s(C_i, C_j) = \frac{1}{2}\left[s(C_{i_1}, C_j) + s(C_{i_2}, C_j)\right] + \frac{1}{4}\left[1 - s(C_{i_1}, C_{i_2})\right]$$

Notemos que una característica de los métodos basados en el centroide y sus variantes es que el valor de similaridad o la distancia asociada con los clusters enlazados puede aumentar o disminuir de una etapa a otra. Por ejemplo, cuando la medida es una distancia, la distancia entre los centroides puede ser menor que la de otro par de centroides unidos en una etapa anterior. Esto puede ocurrir ya que los centroides, en cada etapa, pueden cambiar de lugar. Este problema puede llevar a que el dendrograma resultante sea complicado de interpretar.

Ejemplo 3.5 *Consideremos los siguientes individuos sobre los cuales se han medido dos variables y apliquemos los métodos del centroide ponderado y el de la mediana, empleando para ello la distancia euclídea al cuadrado.*

Individuo	X_1	X_2
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10

Método del Centroide Ponderado.

1. **Nivel 1:**

   La matriz inicial de distancias es

	A	**B**	**C**	**D**	**E**
A	0				
B	325	0			
C	425	200	0		
D	500	125	25	0	
E	50	325	625	650	0

   A la vista de esta matriz se unen los individuos C y D. El centroide del cluster (C, D) es $(30, 12{,}5)$.

2. **Nivel 2:**

   La matriz de distancias en este paso es

	A	**B**	**(C,D)**	**E**
A	0			
B	325	0		
(C,D)	456,25	156,25	0	
E	50	325	631,25	0

   uniéndose en este nivel los individuos A y E. El centroide del cluster (A, E) es $(7{,}5, 7{,}5)$.

3. **Nivel 3:**

   La matriz de distancias en este nivel es

	(A,E)	**B**	**(C,D)**
(A,E)	0		
B	312,5	0	
(C,D)	531,25	156,25	0

   En este nivel se unen los clusters (C, D) y B. El centroide del cluster (B, C, D) es $(26{,}66, 15)$.

4. **Nivel 4:**

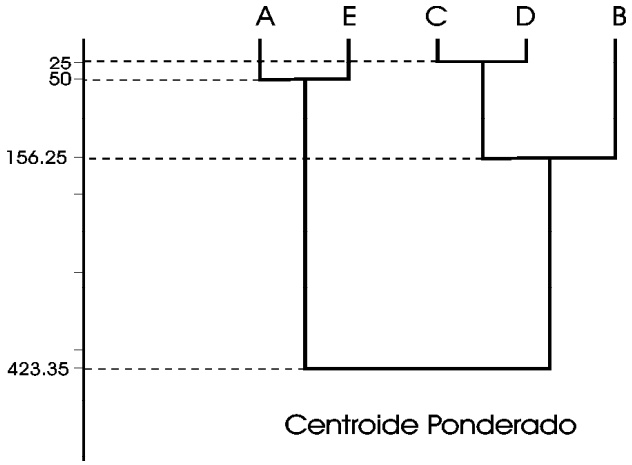
La matriz de distancias en este nivel es

	(A,E)	**(B,C,D)**
(A,E)	0	
(B,C,D)	423,35	0

completándose así la jerarquía. El centroide es el punto $(19, 12)$.

El dendrograma asociado es el de la figura 3.6

Figura 3.6: Método del centroide ponderado



Método de la mediana.

1. **Nivel 1:**

La matriz inicial de distancias es

	A	**B**	**C**	**D**	**E**
A	0				
B	325	0			
C	425	200	0		
D	500	125	25	0	
E	50	325	625	650	0

A la vista de esta matriz se unen los individuos C y D. El centroide del cluster (C, D) *es* $(30, 12{,}5)$.

2. **Nivel 2:**

La matriz de distancias en este paso es

	A	**B**	**(C,D)**	**E**
A	0			
B	325	0		
(C,D)	456,25	156,25	0	
E	50	325	631,25	0

uniéndose en este nivel los individuos A y E. El centroide del cluster (A, E) *es* $(7{,}5, 7{,}5)$.

3. **Nivel 3:**

La matriz de distancias en este nivel es

	(A,E)	**B**	**(C,D)**
(A,E)	0		
B	312,5	0	
(C,D)	531,25	156,25	0

En este nivel se unen los clusters (C, D) *y* B*. El centroide del cluster* (B, C, D) *es* $(25, 16{,}25)$*.*

4. **Nivel 4:**

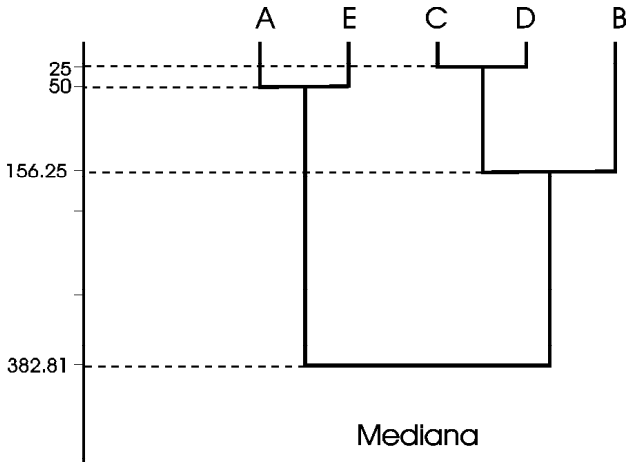
La matriz de distancias en este nivel es

	(A,E)	**(B,C,D)**
(A,E)	0	
(B,C,D)	382,81	0

completándose así la jerarquía. El centroide es el punto $(16{,}25, 11{,}875)$

El dendrograma asociado es el de la figura 3.7

Figura 3.7: Método de la mediana



3.2.6. Método de Ward.

El método de Ward es un procedimiento jerárquico en el cual, en cada etapa, se unen los dos clusters para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de las diferencias, dentro de cada cluster, de cada individuo al centroide del cluster.

Notemos por

- x_{ij}^k al valor de la $j-$ésima variable sobre el $i-$ésimo individuo del $k-$ésimo cluster, suponiendo que dicho cluster posee n_k individuos.
- m^k al centroide del cluster k, con componentes m_j^k.
- E_k a la suma de cuadrados de los errores del cluster k, o sea, la distancia euclídea al cuadrado entre cada individuo del cluster k a su centroide

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n} (x_{ij}^k - m_j^k)^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{n} (x_{ij}^k)^2 - n_k \sum_{j=1}^{n} (m_j^k)^2$$

- E a la suma de cuadrados de los errores para todos los clusters, o sea, si suponemos que hay h clusters

$$E = \sum_{k=1}^{h} E_k$$

El proceso comienza con m clusters, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del cluster y por lo tanto en este primer paso se tendrá $E_k = 0$ para cada cluster y con ello, $E = 0$. El objetivo del método de Ward es encontrar en cada etapa aquellos dos clusters cuya unión proporcione el menor incremento en la suma total de errores, E.

Supongamos ahora que los clusters C_p y C_q se unen resultando un nuevo cluster C_t. Entonces el incremento de E será

$$\Delta E_{pq} = E_t - E_p - E_q =$$

$$= \left[\sum_{i=1}^{n_t}\sum_{j=1}^{n}(x_{ij}^t)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2\right] - \left[\sum_{i=1}^{n_p}\sum_{j=1}^{n}(x_{ij}^p)^2 - n_p\sum_{j=1}^{n}(m_j^p)^2\right] - \left[\sum_{i=1}^{n_q}\sum_{j=1}^{n}(x_{ij}^q)^2 - n_q\sum_{j=1}^{n}(m_j^q)^2\right] =$$

$$= n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2$$

Ahora bien

$$n_t m_j^t = n_p m_j^p + n_q m_j^q$$

de donde

$$n_t^2(m_j^t)^2 = n_p^2(m_j^p)^2 + n_q^2(m_j^q)^2 + 2n_p n_q m_j^p m_j^q$$

y como

$$2m_j^p m_j^q = (m_j^p)^2 + (m_j^q)^2 - (m_j^p - m_j^q)^2$$

se tiene

$$n_t^2(m_j^t)^2 = n_p(n_p + n_q)(m_j^p)^2 + n_q(n_p + n_q)(m_j^q)^2 - n_p n_q(m_j^p - m_j^q)^2$$

Dado que $n_t = n_p + n_q$, dividiendo por n_t^2 se obtiene

$$(m_j^t)^2 = \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

con lo cual se obtiene la siguiente expresión de ΔE_{pq}:

$$n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - n_t\sum_{j=1}^{n}\left[\frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2\right]$$

$$\Delta E_{pq} = n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - m_p\sum_{j=1}^{n}(m_j^p)^2 - n_q\sum_{j=1}^{n}(m_j^q)^2 + \frac{n_p n_q}{n_t}\sum_{j=1}^{n}(m_j^p - m_j^q)^2$$

$$= \frac{n_p n_q}{n_t}\sum_{j=1}^{n}(m_j^p - m_j^q)^2$$

Así el menor incremento de los errores cuadráticos es proporcional a la distancia euclídea al cuadrado de los centroides de los clusters unidos. La suma E es no decreciente y el método, por lo tanto, no presenta los problemas de los métodos del centroide anteriores.

Veamos, para finalizar, cómo se pueden calcular los distintos incrementos a partir de otros calculados con anterioridad.

Sea C_t el cluster resultado de unir C_p y C_q y sea C_r otro cluster distinto a los otros dos. El incremento potencial en E que se produciría con la unión de C_r y C_t es

$$\Delta E_{rt} = \frac{n_r n_t}{n_r + n_t}\sum_{j=1}^{n}(m_j^r - m_j^t)^2$$

Teniendo en cuenta que

$$\begin{aligned} m_j^t = & \quad \frac{n_p m_j^p + n_q m_j^q}{n_t} \\ n_t = & \quad n_p + n_q \end{aligned}$$

y la expresión

$$(m_j^t)^2 = \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

se deduce

$$(m_j^r - m_j^t)^2 = (m_j^r)^2 + (m_j^t)^2 - 2m_j^r m_j^t =$$

$$= (m_j^r)^2 + \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2 - 2m_j^r \frac{n_p m_j^p + n_q m_j^q}{n_t} =$$

$$= \frac{n_p (m_j^r)^2 + n_q (m_j^r)^2}{n_t} + \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 -$$

$$- \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2 - 2m_j^r \frac{n_p m_j^p + n_q m_j^q}{n_t} =$$

$$= \frac{n_p}{n_t}(m_j^r - m_j^p)^2 + \frac{n_q}{n_t}(m_j^r - m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2$$

con lo cual

$$\Delta E_{rt} = \frac{n_r n_t}{n_r + n_t} \sum_{j=1}^{n} (m_j^r - m_j^t)^2 =$$

$$= \frac{n_r n_t}{n_r + n_t} \sum_{j=1}^{n} \left[\frac{n_p}{n_t}(m_j^r - m_j^p)^2 + \frac{n_q}{n_t}(m_j^r - m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2 \right] =$$

$$= \frac{n_r n_p}{n_r + n_t} \sum_{j=1}^{n} (m_j^r - m_j^p)^2 + \frac{n_q n_r}{n_r + n_t} \sum_{j=1}^{n} (m_j^r - m_j^q)^2 - \frac{n_r n_p n_q}{n_t (n_r + n_t)} \sum_{j=1}^{n} (m_j^p - m_j^q)^2 =$$

$$= \frac{1}{n_r + n_t} \sum_{j=1}^{n} \left[n_r n_p (m_j^r - m_j^p)^2 + n_r n_q (m_j^r - m_j^q)^2 - \frac{n_r n_p n_q}{n_p + n_q}(m_j^p - m_j^q)^2 \right] =$$

$$= \frac{1}{n_r + n_t} \left[(n_r + n_p)\Delta E_{rp} + (n_r + n_q)\Delta E_{rq} - n_r \Delta E_{pq} \right]$$

Al igual que en los anteriores métodos del centroide se puede demostrar que la relación anterior se sigue verificando para una distancia que venga definida a partir de una norma que proceda de un producto escalar o que verifique la ley del paralelogramo.

Ejemplo 3.6 *Veamos cómo funciona este procedimiento en el caso de 5 individuos sobre los cuales se miden dos variables. Los datos son los siguientes*

Individuo	X_1	X_2
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10

Nivel 1

En este primer paso hemos de calcular las $\binom{5}{2} = 10$ *posibles combinaciones.*

Partición	*Centroides*	E_k	E	ΔE
$(A,B),C,D,E$	$C_{AB}=(15,12{,}5)$	$E_{AB}=162{,}5$ $E_C=E_D=E_E=0$	162,5	162,5
$(A,C),B,D,E$	$C_{AC}=(20,7{,}5)$	$E_{AC}=212{,}5$ $E_B=E_D=E_E=0$	212,5	212,5
$(A,D),B,C,E$	$C_{AD}=(20,10)$	$E_{AD}=250$ $E_B=E_C=E_E=0$	250	250
$(A,E),B,C,D$	$C_{AE}=(7{,}5,7{,}5)$	$E_{AE}=25$ $E_B=E_C=E_D=0$	25	25
$(B,C),A,D,E$	$C_{BC}=(25,15)$	$E_{BC}=100$ $E_A=E_D=E_E=0$	100	100
$(B,D),A,C,E$	$C_{BD}=(25,17{,}5)$	$E_{BD}=62{,}5$ $E_A=E_C=E_E=0$	62,5	62,5
$(B,E),A,C,D$	$C_{BE}=(12{,}5,15)$	$E_{BE}=162{,}5$ $E_A=E_C=E_D=0$	162,5	162,5
$(C,D),A,B,E$	$C_{CD}=(30,12{,}5)$	$E_{CD}=12{,}5$ $E_A=E_B=E_E=0$	12,5	12,5
$(C,E),A,B,D$	$C_{CE}=(17{,}5;10)$	$E_{CE}=312{,}5$ $E_A=E_B=E_D=0$	312,5	312,5
$(D,E),A,B,C$	$C_{DE}=(17{,}5;12{,}5)$	$E_{DE}=325$ $E_A=E_B=E_C=0$	325	325

de donde se deduce que en esta etapa se unen los elementos C y D. La configuración actual es $(C,D),A,B,E$.

Nivel 2

A partir de la configuración actual tomamos las $\binom{4}{2}=6$ combinaciones posibles.

Partición	*Centroides*	E_k	E	ΔE
$(A,C,D),B,E$	$C_{ACD}=(23{,}33,10)$	$E_{ACD}=316{,}66$ $E_B=E_E=0$	316,66	304,16
$(B,C,D),A,E$	$C_{BCD}=(26{,}66,15)$	$E_{BCD}=116{,}66$ $E_A=E_E=0$	116,66	104,16
$(C,D,E),A,B$	$C_{CDE}=(21{,}66,11{,}66)$	$E_{CDE}=433{,}33$ $E_A=E_B=0$	433,33	420,83
$(A,B),(C,D),E$	$C_{AB}=(15,12{,}5)$ $C_{CD}=(30,12{,}5)$	$E_{AB}=162{,}5$ $E_{CD}=12{,}5$ $E_E=0$	175	162,5
$(A,E),(C,D),B$	$C_{AE}=(7{,}5,7{,}5)$ $C_{CD}=(30,12{,}5)$	$E_{AE}=25$ $E_{CD}=12{,}5$ $E_B=0$	37,5	25
$(B,E),(C,D),A$	$C_{BE}=(12{,}5,15)$ $C_{CD}=(30,12{,}5)$	$E_{BE}=162{,}5$ $E_{CD}=12{,}5$ $E_A=0$	175	162,5

de donde se deduce que en esta etapa se unen los elementos A y E. La configuración actual es $(A,E),(C,D),B$.

Paso 3

A partir de la configuración actual tomamos las $\binom{3}{2}=3$ combinaciones posibles.

Partición	*Centroides*	E_k	E	ΔE
$(A,C,D,E),B$	$C_{ACDE}=(18{,}75,10)$	$E_{ACDE}=568{,}75$ $E_B=0$	568,75	531,25
$(A,B,E),(C,D)$	$C_{ABE}=(11{,}66,11{,}66)$ $C_{CD}=(30,12{,}5)$	$E_{ABE}=233{,}33$ $E_{CD}=12{,}5$	245,8	208,3
$(A,E),(B,C,D)$	$C_{AE}=(7{,}5,7{,}5)$ $C_{BCD}=(26{,}66,15)$	$E_{AE}=25$ $E_{BCD}=116{,}66$	141,66	104,16

de donde se deduce que en esta etapa se unen los clusters B y (C,D). La configuración actual es $(A,E),(B,C,D)$.

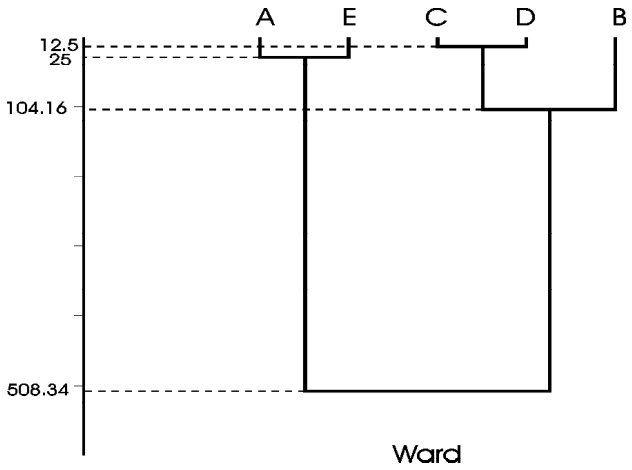
Paso 4

Evidentemente en este paso se unirán los dos clusters existentes. Los valores del centroide y de los incrementos de las distancias serán los siguientes

Partición	*Centroide*	E	ΔE
(A, B, C, D, E)	$C_{ABCDE} = (19, 12)$	650	508,34

El dendrograma asociado es el de la figura 3.8

Figura 3.8: Método de Ward



3.3. Fórmula de recurrencia de Lance y Williams.

A continuación vamos a exponer una expresión debida a Lance y Williams en 1967 que intenta aglutinar todos los métodos anteriores bajo una misma fórmula. Concretamente la expresión que dedujeron dichos autores proporciona la distancia entre un grupo K y otro grupo (I, J) formado en una etapa anterior por la fusión de dos grupos. Obviamente dicha expresión tiene importantes aplicaciones desde el punto de vista computacional ya que permite una reducción considerable en los cálculos.

La fórmula en cuestión es la siguiente

$$d(K,(I,J)) = \alpha_I d(K,I) + \alpha_J d(K,J) + \beta d(I,J) + \gamma |d(K,I) - d(K,J)|$$

De esta manera el cálculo de las distancias entre grupos usadas por las técnicas jerárquicas descritas anteriormente son casos particulares de la expresión anterior, para una elección conveniente de los parámetros α_I, α_J, β y γ. Algunos de estos coeficientes han sido ya deducidos en la descripción de los métodos anteriores (métodos del promedio ponderado y no ponderado, método del centroide, método de la mediana y método de Ward).

Veamos ahora cómo el método del *amalgamamiento simple* y el del *amalgamamiento completo* pueden ser también englobados bajo esta filosofía.

Amalgamamiento simple

Supongamos que en una etapa se dispone de un cluster C_j y de otro C_i que es fruto de la unión de otros dos clusters, C_{i_1} y C_{i_2} en una etapa anterior. El método del amalgamamiento simple determina que la distancia entre ambos clusters se establece como la menor distancia existente entre los elementos de ambos clusters; evidentemente, al estar constituido el cluster C_i por otros dos clusters C_{i_1} y C_{i_2}, dicho criterio equivale a calcular el mínimo de las distancias entre el cluster C_j y C_{i_1} y entre C_j y C_{i_2}. Teniendo en cuenta la siguiente igualdad (de fácil comprobación)

$$\text{Min}(a,b) = \frac{1}{2}(a+b) - \frac{1}{2}|a-b|$$

se tiene

$$d(C_j, C_i) = \text{Min}\left\{d(C_j, C_{i_1}), d(C_j, C_{i_2})\right\} =$$

$$= \frac{1}{2}d(C_j, C_{i_1}) + \frac{1}{2}d(C_j, C_{i_2}) - \frac{1}{2}|d(C_j, C_{i_1}) - d(C_j, C_{i_2})|$$

que corresponde a la expresión anterior con

$$\alpha_I = \alpha_J = \frac{1}{2} \;\; ; \;\; \beta = 0 \;\; ; \;\; \gamma = -\frac{1}{2}$$

Amalgamamiento completo

En las mismas hipótesis que en el caso anterior y usando la expresión

$$\text{Max}(a, b) = \frac{1}{2}(a + b) + \frac{1}{2}|a - b|$$

se tiene para el método del *amalgamamiento completo*

$$d(C_j, C_i) = \text{Max}\left\{d(C_j, C_{i_1}), d(C_j, C_{i_2})\right\} =$$

$$= \frac{1}{2}d(C_j, C_{i_1}) + \frac{1}{2}d(C_j, C_{i_2}) + \frac{1}{2}|d(C_j, C_{i_1}) - d(C_j, C_{i_2})|$$

que corresponde a la fórmula de Lance y Williams con

$$\alpha_I = \alpha_J = \frac{1}{2} \;\; ; \;\; \beta = 0 \;\; ; \;\; \gamma = \frac{1}{2}$$

Extrayendo los resultados obtenidos en apartados anteriores para otros procedimientos se puede comprobar la validez de la fórmula de recurrencia para dichos parámetros. Concretamente:

1. **Método del promedio no ponderado**

$$\alpha_I = \alpha_J = \frac{1}{2} \;\; ; \;\; \beta = \gamma = 0$$

2. **Método del promedio ponderado**

$$\alpha_I = \frac{n_{i_1}}{n_{i_1} + n_{i_2}} \;\; ; \;\; \alpha_J = \frac{n_{i_2}}{n_{i_1} + n_{i_2}} \;\; ; \;\; \beta = \gamma = 0$$

3. **Método del centroide**

   Para la distancia euclídea al cuadrado se tiene

$$\alpha_I = \frac{n_{i_1}}{n_{i_1} + n_{i_2}} \;\; ; \;\; \alpha_J = \frac{n_{i_2}}{n_{i_1} + n_{i_2}} \;\; ; \;\; \beta = -\alpha_I\alpha_J \;\; ; \;\; \gamma = 0$$

4. **Método de la mediana**

$$\alpha_I = \alpha_J = \frac{1}{2} \;\; ; \;\; \beta = -\frac{1}{4} \;\; ; \;\; \gamma = 0$$

5. **Método de Ward**

   Para la distancia euclídea al cuadrado se tiene

$$\alpha_I = \frac{n_{i_1} + n_j}{n_{i_1} + n_{i_2} + n_j} \;\; ; \;\; \alpha_J = \frac{n_{i_2} + n_j}{n_{i_1} + n_{i_2} + n_j} \;\; ; \;\; \beta = -\frac{n_j}{n_{i_1} + n_{i_2} + n_j} \;\; ; \;\; \gamma = 0$$

3.4. Métodos Jerárquicos Disociativos.

Como se comentó en la introducción de este capítulo, los métodos disociativos, constituyen el proceso inverso a los aglomerativos. Comienzan con un conglomerado que engloba a todos los casos tratados y, a partir de este grupo inicial, a través de sucesivas divisiones, se van formando grupos cada vez menores. Al final del proceso se tienen tantas agrupaciones como casos han sido tratados.

En cuanto a la clasificación de estos métodos se puede decir que la filosofía de los métodos aglomerativos puede mantenerse para este otro tipo de procedimientos en lo que concierne a la forma de calcular la distancia entre los grupos, si bien, como es lógico, al partir de un grupo único que hay que subdividir, se seguirá la estrategia de maximizar las distancias, o minimizar las similaridades, puesto que buscamos ahora los individuos menos similares para separarlos del resto del conglomerado.

Esta clase de métodos son esencialmente de dos tipos:

1. Monotéticos, los cuales dividen los datos sobre la base de un solo atributo y suelen emplearse cuando los datos son de tipo binario.

2. Politéticos, cuyas divisiones se basan en los valores tomados por todas las variables.

Esta clase de procedimientos es bastante menos popular que los ascendentes por lo que la literatura sobre ellos no es muy extensa. Una cuestión importante que puede surgir en su desarrollo es el hecho de cuándo un cluster determinado debe dejar de dividirse para proceder con la división de otro conglomerado distinto. Dicha cuestión puede resolverse con la siguiente variante expuesta por MacNaughton-Smith en 1964 y que está concebida para aquellas medidas de asociación que sean positivas.

Dicho procedimiento comienza con la eliminación del grupo principal de aquel individuo cuya distancia sea mayor, o cuya similaridad sea menor, al cluster formado por los restantes individuos, tomando como base para calcular dichas distancias o similaridades cualquiera de los procedimientos anteriormente descritos en los métodos ascendentes. Así se tiene un cluster unitario y otro formado por los restantes individuos.

A continuación se añadirá al cluster unitario aquel elemento cuya distancia (similaridad) total al resto de los elementos que componen su actual cluster menos la distancia (similaridad) al cluster anteriormente formado sea máxima (mínima). Cuando esta diferencia sea negativa dicho elemento no se añade y se repite el proceso sobre los dos subgrupos.

Ejemplo 3.7 *Retomemos la matriz de distancias del ejemplo 3.1*

El método de cálculo de las distancias será la del método del amalgamamiento simple. *(Se propone como ejercicio el empleo de los otros tipos de estrategia).*

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	2,15	0					
C	0,7	1,53	0				
D	1,07	1,14	0,43	0			
E	0,85	1,38	0,21	0,29	0		
F	1,16	1,01	0,55	0,22	0,41	0	
G	1,56	2,83	1,86	2,04	2,02	2,05	0

Paso 1

Las distancias de cada individuo al cluster formado por el resto es

A	0,7
B	1,01
C	0,21
D	0,22
E	0,21
F	0,22
G	1,56

por lo que el individuo empleado para comenzar la división será el individuo etiquetado **G** *(notemos que ahora el criterio que se sigue es maximizar la distancia). Tenemos con ello dos clusters,* (G) *y* (A, B, C, D, E, F).

Paso 2

A continuación calculamos la distancia de cada individuo del cluster principal al resto, la distancia de cada individuo de dicho grupo al nuevo cluster formado así como la diferencia entre ambas.

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
A	0,7	1,56	−0,86
B	1,01	2,83	−1,82
C	0,21	1,86	−1,65
D	0,22	2,04	−1,82
E	0,21	2,02	−1,81
F	0,22	2,05	−1,83

A la vista de estos resultados es obvio que ningún elemento se añadirá al cluster anterior, por lo que procede comenzar con la división del grupo principal, empezando por el individuo **B**. *Tenemos así la división* (G), (B) (A, C, D, E, F).

Paso 3

Volvemos a calcular la distancia entre cada individuo del cluster (A, C, D, E, F) *así como la distancia de cada individuo de dicho grupo al nuevo cluster formado y la diferencia entre ambas.*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
A	0,7	2,15	−1,45
C	0,21	1,53	−1,32
D	0,22	1,14	−0,92
E	0,21	1,38	−1,17
F	0,22	1,01	−0,79

de donde se deduce que ningún individuo se añadirá al nuevo cluster formado. Ahora se empezará a dividir el cluster (A, C, D, E, F) *por el individuo* **A**.

Paso 4

Calculamos la distancia entre cada individuo del cluster (C, D, E, F) *así como la distancia de cada individuo de dicho grupo al nuevo cluster formado y la diferencia entre ambas.*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
C	0,21	0,7	−0,49
D	0,22	1,07	−0,85
E	0,21	0,85	−0,64
F	0,22	1,16	−0,94

Ningún elemento se añadirá al cluster formado por el individuo **A**. *Elegimos ahora el individuo* **D** *(también podíamos haber elegido el* **F***).*

Paso 5

Calculamos la distancia entre cada individuo del cluster (C, E, F) *así como la distancia de cada individuo de dicho grupo al nuevo cluster formado y la diferencia entre ambas.*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
C	0,21	0,43	−0,22
E	0,21	0,29	−0,08
F	0,41	0,22	0,19

A la vista del resultado anterior se tiene que el individuo **F** *se suma al individuo* **D**. *Vemos si algún otro individuo se une*

Indiv.	*Distancia en el grupo principal*	*Distancia al nuevo cluster*	*Diferencia*
C	0,21	0,43	−0,22
E	0,21	0,29	−0,08

con lo cual no se añade ningún individuo.

Paso 6

El proceso se seguiría ahora descomponiendo los dos clusters que quedan, a saber, (D, F) *y* (C, E)*, empezando con el primero de ellos pues es el que más distancia presenta entre sus elementos.*

Las técnicas monotéticas son generalmente empleadas cuando los datos son de tipo binario. Ahora la división se inicia en aquellos individuos que poseen y aquellos que no poseen algún atributo específico. Teniendo en cuenta este criterio, para un conjunto de datos con m variables binarias hay m divisiones potenciales del conjunto inicial, $m-1$ para cada uno de los dos subgrupos formados y así sucesivamente; de ello se deduce que hay que determinar algún criterio para elegir la variable sobre la cual se va a proceder a la división.

El criterio que suele ser más usual es el basado en los estadísticos del tipo χ^2 obtenidos a partir de la tabla de doble entrada para cada par de variables

$$\chi^2_{jk} = \frac{(ad-bc)^2 N}{(a+b)(a+c)(b+d)(c+d)}$$

y tomar la variable k tal que $\sum_{j \neq k} \chi^2_{jk}$ sea máximo.

Otros criterios alternativos pueden ser

$$\text{Max} \sum \sqrt{\chi^2_{jk}}$$

$$\text{Max} \sum |ad-bc|$$

$$\text{Max} \sum (ad-bc)^2$$

Por ejemplo consideremos el siguiente ejemplo en el cual se tienen 5 individuos sobre los cuales se miden tres variables de tipo binario

X_1	X_2	X_3
0	1	1
1	1	0
1	1	1
1	1	0
0	0	1

Calculemos primero los estadísticos χ^2 para cada par de variables. Por ejemplo, para las variables X_1 y X_2 se tiene

$X_2 \backslash X_1$	1	0	Total
1	3	1	4
0	0	1	1
Total	3	2	5

de donde

$$\chi^2_{12} = \frac{45}{24} = 1{,}875$$

Asimismo $\chi^2_{13} = \frac{80}{36} = 2{,}22$ y $\chi^2_{23} = \frac{20}{24} = 0{,}83$. Ahora, aplicando el criterio $\text{Max} \sum_{j \neq k} \chi^2_{jk}$, se tiene

$$\chi^2_{12} + \chi^2_{13} = 4{,}09$$
$$\chi^2_{12} + \chi^2_{23} = 2{,}7$$
$$\chi^2_{13} + \chi^2_{23} = 3{,}05$$

de donde la división se basará en la determinación de quien posee la característica asociada a la variable X_1 y quien no, obteniéndose así los dos clusters (I_2, I_3, I_4) y (I_1, I_5). De forma sucesiva se seguiría aplicando este criterio a ambos subgrupos.

3.5. La matriz cofenética. Coeficiente de correlación cofenético.

Los métodos jerárquicos imponen una estructura sobre los datos y es necesario con frecuencia considerar si es aceptable o si se introducen distorsiones inaceptables en las relaciones originales. El método más usado para verificar este hecho, o sea, para ver la relación entre el dendrograma y la matriz de proximidades original, es el coeficiente de correlación cofenético, el cual es simplemente la correlación entre los $\frac{n(n-1)}{2}$ elementos de la parte superior de la matriz de proximidades observada y los correspondientes en la llamada matriz cofenética, C, cuyos elementos, c_{ij}, se definen como aquellos que determinan la proximidad entre los elementos i y j cuando éstos se unen en el mismo cluster.

Así, si tras el empleo de varios procedimientos cluster distintos, éstos conducen a soluciones parecidas, surge la pregunta de qué método elegiremos como definitivo. La respuesta la da el coeficiente cofenético, ya que aquel método que tenga un coeficiente cofenético más elevado será aquel que presente una menor distorsión en las relaciones originales existentes entre los elementos en estudio.

Ejemplo 3.8 *Calculemos las matrices cofenéticas y los coeficientes de correlación cofenéticos asociados a los ejemplos 3.1 a 3.4*

1. **Método del amalgamamiento simple**

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	1,01	0					
C	0,7	1,01	0				
D	0,7	1,01	0,29	0			
E	0,7	1,01	0,21	0,29	0		
F	0,7	1,01	0,29	0,22	0,29	0	
G	1,56	1,56	1,56	1,56	1,56	1,56	0

siendo el coeficiente de correlación cofenético 0.911438774

2. **Método del amalgamamiento completo**

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	2,83	0					
C	1,16	2,83	0				
D	1,16	2,83	0,55	0			
E	1,16	2,83	0,21	0,55	0		
F	1,16	2,83	0,55	0,22	0,55	0	
G	2,05	2,83	2,05	2,05	2,05	2,05	0

siendo el coeficiente de correlación cofenético 0.788405653

3. **Método de la distancia promedio no ponderada**

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	1,7075	0					
C	0,945	1,7075	0				
D	0,945	1,7075	0,41	0			
E	0,945	1,7075	0,21	0,41	0		
F	0,945	1,7075	0,41	0,22	0,41	0	
G	2,303125	2,303125	2,303125	2,303125	2,303125	2,303125	0

siendo el coeficiente de correlación cofenético 0.911167777

4. **Método de la distancia promedio ponderada**

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0						
B	1,442	0					
C	0,945	1,442	0				
D	0,945	1,442	0,42	0			
E	0,945	1,442	0,21	0,42	0		
F	0,945	1,442	0,42	0,22	0,42	0	
G	2,06	2,06	2,06	2,06	2,06	2,06	0

siendo el coeficiente de correlación cofenético 0.911359728

3.6. El problema del número de clusters a determinar.

Con frecuencia, cuando se emplean técnicas clusters jerárquicas, el investigador no está interesado en la jerarquía completa sino en un subconjunto de particiones obtenidas a partir de ella. Las particiones se obtienen cortando el dendrograma o seleccionando una de las soluciones en la sucesión encajada de clusters que comprende la jerarquía.

Desafortunadamente este paso fundamental está entre los problemas que todavía no están totalmente resueltos. Entre las razones más importantes que se pueden citar para que dicho problema siga siendo un campo abierto están las siguientes:

1. **La inexistencia de una hipótesis nula apropiada.**

   En efecto, la dificultad para crear una hipótesis nula operativa radica en la falta de una definición clara y comprensiva de lo que significa *no estructura* en un conjunto de datos. El concepto de *no estructura* (que podía ser una posible hipótesis nula) está bastante lejos de ser clara, lo cual conlleva a no saber qué tipos de contrastes hay que desarrollar para determinar si una determinada estructura está presente o no en el conjunto de datos. Dubes y Jain (1980) comentan sobre este hecho lo siguiente:

   > *... el rechazo de la hipótesis nula no es significativo porque no han sido desarrolladas hipótesis alternativas significativas; todavía no existe una definición útil y práctica de estructura cluster, matemáticamente hablando.*

2. **La naturaleza compleja de las distribuciones muestrales multivariantes.**

   Igualmente intratable es el problema de la mixtura de las distribuciones muestrales multivariantes en el análisis de datos reales. Aunque son muchos los aspectos conocidos y desarrollados acerca de la distribución normal multivariante, no es ni esperable ni razonable que los datos que se manejen en estos estudios obedezcan a dicha ley, sino que existirán mixturas de diversas distribuciones muestrales que pueden ser incluso desconocidas.

Las soluciones propuestas a estas cuestiones han sido múltiples. En algunos campos de aplicación, como puede ser algunos tipos de investigaciones en las ciencias biológicas, el problema de determinar el número de clusters no es un tema que parezca excesivamente importante ya que el objetivo puede ser simplemente explorar el patrón general de las relaciones existentes entre los individuos objeto de estudio, lo cual puede ser observado a partir del dendrograma. Sin embargo hay campos de aplicación en los cuales se pretende ir más lejos en el estudio y obtener una clasificación de los individuos lo más realista posible, lo cual conlleva tener que estudiar con más énfasis el problema del número de clusters a determinar. Esta cuestión ha motivado la aparición de múltiples *reglas*. Algunas de estas reglas son simples métodos heurísticos, otras están basadas en contrastes de hipótesis formales, los cuales han sido desarrollados al amparo de la hipótesis de la existencia de una determinada distribución muestral (casi siempre la normal multivariante), mientras que otros son procedimientos asimismo heurísticos pero que extraen la filosofía de los contrastes existentes en poblaciones normales. A continuación vamos a citar algunas de estas reglas, si bien hay que decir que son muchísimos los procedimientos que en los últimos años han sido desarrollados, con frecuencia orientados a técnicas particulares.

• La primera técnica que podemos citar se basa simplemente en cortar el dendrograma de forma subjetiva tras visualizarlo. Obviamente este procedimiento no es nada satisfactorio puesto que está generalmente sesgado por la opinión que el investigador posee sobre sus datos.

• Un método más formal, pero asimismo heurístico, se basa en representar en una gráfica el número de clusters que se observan en los distintos niveles del dendrograma frente a los niveles de fusión a los que los

clusters se unen en cada nivel. La presencia de una pendiente poco pronunciada sugiere que la siguiente unión de clusters no aporta apenas información adicional sobre la aportada en el nivel anterior. Este método, por lo tanto, se basa en la existencia de *pequeños* saltos o discontinuidades en los niveles de fusión.

• Mojena (1977) siguió con la idea de estudiar los saltos relativos en los valores de fusión y sugirió otro procedimiento heurístico bastante divulgado y que ha sido fuente de bastantes investigaciones posteriores. En su método se compara el valor de fusión de cada etapa con el promedio de los valores de fusión sumado con el producto de una cierta constante por la cuasidesviación típica de los valores de fusión. Cuando un valor de fusión supera dicha cantidad se concluye que el nivel precedente es el que origina la solución óptima. Mojena sugirió que el valor de la constante debía de estar comprendido en el rango de 2.75 a 3.50 si bien Milligan, en 1985, tras una detallada investigación de valores en función del número de clusters, establece que el valor óptimo para dicha constante debe ser 1.25.

• Beale en (1969) propuso el uso de un contraste basado en la distribución F de Snedecor para contrastar la hipótesis de la existencia de c_2 clusters frente a la existencia de c_1 clusters, siendo $c_2 > c_1$. Para ello se consideran la suma, para cada partición, de las desviaciones cuadráticas medias de los elementos de cada cluster a su centroide, llamémoslas DC_1 y DC_2:

$$DC_1 = \frac{1}{n - c_1} \sum_{i=1}^{c_1} \sum_{j=1}^{n_i} ||x_{ij} - \overline{x}_i||^2$$

$$DC_2 = \frac{1}{n - c_2} \sum_{i=1}^{c_2} \sum_{j=1}^{n_i} ||x_{ij} - \overline{x}_i||^2$$

donde se ha supuesto que el cluster $i-$ésimo posee n_i elementos y n es el total de la muestra. El estadístico considerado es

$$F(p(c_2 - c_1), p(n - c_2)) = \frac{\dfrac{DC_1 - DC_2}{DC_2}}{\left[\left(\dfrac{n - c_1}{n - c_2}\right)\left(\dfrac{c_2}{c_1}\right)^{\frac{2}{p}} - 1\right]}$$

Un resultado significativo indica que la división en c_2 clusters representa una mejoría frente a la división en c_1 clusters. Notemos que este contraste no impone ninguna distribución concreta de la muestra.

Los siguientes métodos que vamos a comentar ahora proceden en su mayoría de la abstracción de procedimientos inherentes en su mayoría al análisis multivariante paramétrico. Para su desarrollo, definimos las siguientes matrices:

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})(x_{ij} - \overline{x})'$$

$$W = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)(x_{ij} - \overline{x}_i)'$$

$$B = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'$$

Estas matrices representan, respectivamente, la dispersión total de todos los individuos respecto de su centroide, la suma de las dispersiones en cada grupo (desviación intra clusters) y la dispersión entre los grupos (desviación entre clusters). Asimismo k representa el número total de clusters y n es el tamaño total de la muestra ($n = n_1 + \cdots + n_k$).

Se puede comprobar que se cumple la igualdad $T = W + B$. Dicha igualdad es la extensión al caso multivariante de la conocida descomposición de la variabilidad del análisis de la varianza de una vía. Para fijar ideas y particularizando al caso unidimensional, es obvio que en tales circunstancias un criterio lógico para determinar el número de clusters sería elegir aquella partición que tuviera el menor valor en la desviación intra-clusters o, equivalentemente, el mayor valor en la desviación entre-clusters.

Siguiendo con esta idea se puede extender dicha situación al caso multivariante, si bien el empleo de las matrices antes reseñadas no hace tan inmediata dicha extensión. Por ello surgen diversos procedimientos, entre los cuales podemos citar los siguientes:

1. **Minimización de la traza de W.**

   Esta es la extensión más inmediata al criterio anteriormente comentado para el caso unidimensional. Evidentemente esto es equivalente a minimizar la suma de los cuadrados de las distancias euclídeas entre cada individuo a la media del cluster al que ha sido asignado.

   Hay que hacer notar que este criterio está implícito en diversos métodos no jerárquicos que serán descritos en el capítulo siguiente, como el de Forgy, Jancey y el de las k-medias, así como, dentro de los métodos jerárquicos, el de Ward.

   Notemos asimismo que como $T = W + B$, entonces $\mathrm{tr}[T] = \mathrm{tr}[W] + \mathrm{tr}[B]$, por lo que minimizar la traza de W equivale a maximizar la traza de B ya que, sea cual sea la configuración de clusters que se establezca, la matriz T no varía y, por tanto, tampoco su traza.

2. **Minimización de $k^2|W|$.**

   Marriot en 1971 sugiere el empleo de $k^2|W|$, tomándose el valor de k tal que haga esa cantidad mínimo.

3. **Minimización del determinante de W.**

   En el análisis de la varianza multivariante de una vía (MANOVA) son diversos los criterios empleados basados en la distribución de la razón de verosimilitudes. Entre ellos destaca el criterio de Wilks, el cual considera el cociente

   $$\Lambda = \frac{|W|}{|T|} = \frac{|W|}{|W+B|}$$

   rechazándose la hipótesis nula de igualdad de las medias poblacionales si ese cociente es menor que un valor predeterminado o, lo que es equivalente, si el cociente

   $$\frac{|T|}{|W|}$$

   es mayor que un determinado valor.

   Es evidente que en nuestro ambiente no podemos aplicar este contraste ya que carecemos de las hipótesis de normalidad multivariante, pero se puede *abstraer* la filosofía de dicho contraste y aplicarlo para nuestros propósitos, lo cual no deja de ser un método puramente heurístico. Así pues y puesto que para todas las particiones de los individuos en k grupos la matriz T permanece constante, Friedman y Rubin sugirieron en 1967 la maximización de dicho cociente, lo cual equivale a la minimización de $|W|$.

4. **Maximización de la traza de BW^{-1}.**

   Siguiendo con la misma idea anterior, otro de los criterios que se pueden aplicar en el análisis de la varianza multivariante de una vía es el debido a Lawley y Hotelling, quienes proponen el empleo del estadístico

   $$\mathrm{tr}[BW^{-1}]$$

   siendo rechazada la hipótesis nula cuando dicha traza supere un cierto valor impuesto de antemano.

   En nuestro caso, y siempre abstrayendo la filosofía del criterio expuesto, debemos seleccionar aquella partición que produzca la maximización de esa traza.

5. Por otro lado, Calinski y Harabasz (1974) proponen el estadístico

   $$C = \frac{\dfrac{\mathrm{tr}[B]}{k-1}}{\dfrac{\mathrm{tr}[W]}{n-k}}$$

   tomando como número óptimo de clusters aquel que produzca el mayor valor de C.